Project funded by the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme FP7/2007-2013/ under REA grant agreement no. 317471.

Project reference: 317471

Project full title: EXPloiting Empirical appRoaches to Translation

D4.2: Terminology and Ontology

Authors: Liling Tan (USAAR)

Contributors: Josef van Genabith (USAAR), Marcos Zampieri (USAAR), Anne Schumann (USAAR), Jon Dehdari (USAAR), Santanu Pal (USAAR), Rohit Gupta (UoW), Hanna Bechara (UoW)

***Document Number*:** EXPERT_D4.2_20150508

***Distribution Level*:** Public

***Contractual Date of Delivery*:** 30.10.14

***Actual Date of Delivery*:** 08.05.15

Contributing to the Deliverable: WP4

WP Task Responsible: USAAR

EC Project Officer: Concepcion Perez-Camaras

Contents

1. Introduction

This document reports the advancement in terminology extraction and ontology induction research carried out in Work Package 4.2 (WP4.2) by Early Stage Researcher 5 (ESR5) under the EXPERT (EXPloiting Empirical appRoaches to Translation) project. The report is split into two parts and an overall conclusion; the first part will describe the experiments on term extraction (Sections 2-4) and the second on ontology induction (Sections 5-7).

2. Terminology

This section provides the terminology portion of the deliverable. We will (i) define the notion of term and terminology, (ii) give a brief survey of term extraction techniques, (iii) introduce our novel approach to terminology extraction and (iv) our preliminary results, and (v) conclude the terminology section of the deliverable.

A ***term*** is the designation of a defined concept in a special language by a linguistic expression; a term may consist of one or more words. A ***terminology*** refers to the set of terms representing the system of concepts of a particular subject field (ISO 1087). The International Organization of Standardization (ISO) history of terminology traces back to Wüster's (1969) seminal article on *Die vier Dimensionen der Terminologiearbeit*[1] which the ISO Technical Committee 37 (ISO/TC 37) builds upon in providing the common standards related to terminology work.

A later formulation states that a term is any conventional symbol representing a concept defined in a subject field; a terminology is the aggregate of terms, which represent the system of concepts of an individual subject field (Felber, 1984). The core characteristic of a term is defined as ***termhood***, i.e. the degree to which a linguistic unit is related to a domain-specific context (Kageura and Umino, 1996). In the case of multi-token terms, additional substantiation is necessary to check its ***unithood***, i.e. the degree of strength or stability of syntagmatic combinations and collocations (Kageura and Umino, 1996).

Single token terms can be perceived as a specialized vocabulary that is used specifically in a domain. The surface word representing the single token term is often polysemous and the usage of the term within a specialized domain may narrow down the set of possible senses or single out a disambiguated sense of the word. For instance, the term "*classifier*" can refer to (i) a morpheme used to indicate the semantic class to which the counted item belongs or (ii) a pre-trained model to identify/distinguish different classes within a dataset. The first definition is mainly used within linguistic research, the second within the machine learning

[1] *The Four Dimensions of Terminological Work.*

domain. However, when classifier is used in computational linguistics, its usage is ambiguous. However, the latter definition of classifier tends to be used more often than the former.

In English, terms are more often multi-word expressions (MWE), primarily nominal phrases, made up of a head noun and its complement adjective(s), prepositional clause(s), or compounding noun(s). Commonly, a complex term can be analysed in terms of a head with one or more modifiers (Hippisley et al. 2005).

3. A Survey of Term Extraction

The three main properties of a term can be classified as linguistic, statistical and distributional. Term extraction algorithms may exploit one or a combination of properties and we note that the stratification between the properties is not always clear. Additionally, the termhood properties across languages vary, making bilingual term extraction more challenging.

3.1. Linguistic Properties

The linguistic properties of a term can be characterized by its syntactic context[2]. For instance, Justeson and Katz (1995) and Daille (2000) used different part-of-speech (POS) patterns to extract nominal phrasal terms[3]:

EN: $((Adj|Noun)+|((Adj|Noun)*(NounPrep)?)(Adj|Noun)*)Noun$
FR: $Noun_1(Adj|(Prep(Det)?)?Noun_2|V_{Inf})$

In the case of English, the compulsory head noun is in the final position preceded by its modifiers whereas in French, it is in the first position followed by its modifiers. The multi-word nature of Romance languages produces more terminological phrases whereas for Germanic languages, the compounding nature of nouns derives more single token lexicalized terms. For example, an equivalent POS pattern for German would have to be replaced by a combination of POS and morphemic pattern:

DE: $((AC|NC)+|((AC|NC)*(NounPrep)?)(AC|NC)*)Noun$

Similar to the $(Adj|Noun)$ pattern in English, the $(AC|NC)$ is a combination of $(Adj(Con)|Noun(Con))$ where an optional connective morpheme is necessary to join the adjacent adjectives/nouns.

These linguistics patterns are usually used as filters to generate a list of multi-word terms of high unithood followed by further statistical measures to re-rank or

[2] Often defined as POS pattern or sub-tree parses.
[3] The POS patterns are encoded in regex automata.

reduce the list (e.g. Bourigault et al. 1996). Frantzi et al. (1998) differentiated two types of filters, viz. a close filter is strict about which strings it permits and an open filter allows more strings in the POS patterns. For example, the English pattern is an open filter that allows a wider range of multi-word term candidates than simply using a /Noun+/ that only allows delexicalized compounding nouns.

Although state-of-art term extraction systems do not solely rely on linguistics patterns, the pattern templates are used as filters to remove candidate terms from the system output (e.g. Milios et al. 2003; Guinovart and Simões 2009).

3.2. Statistical Properties

The basis of all statistical properties in multi-word term extraction relies on the frequency of a token or an n-gram in a corpus. Frequency counts are combined to compute co-occurrence measures (aka. word/lexical association measures) that quantify the probabilistic occurrence of a word with its neighbouring words. Co-occurrence measures are used to estimate the propensity for words occurring together.

Psycholinguistic evidences show that word association norms can be measured as a subject's responses to words when preceded by associated words (Palermo and Jenkins, 1964) and humans respond quicker in the case of highly associated words within the same domain (Church and Hanks, 1990).

Common co-occurrence measures, e.g. Dice coefficient, Mutual Information (MI), Pointwise Mutual Information (PMI), Log-Likelihood Ratio (LLR) and Phi-square (Φ2) rely on two main variations of three types of frequency information; (i) the frequency of a word occurring in the corpus, (ii) the joint frequency of a word occurring with another word, (iii) the total number of words in the corpus. Formally we describe them as follows:

Let f_i be the frequency of the occurrence of a word, i

Let f_{ij} be the frequency of the word i and j occurring simultaneously

Let $f_{ij'}$ be the frequency of the word i occurring in the absence of j

Let $f_{i'j}$ be the frequency of the word j occurring in the absence of i

Let $f_{i'j'}$ be the frequency of both words i and j not occurring

Let N be the size of the corpus

We further simplify the notion by having $a = f_{ij}$, $b = f_{ij'}$, $c = f_{i'j}$ and $d = f_{i'j'}$

These basic statistical properties of word co-occurrence are combined in various ways to form more complex co-occurrence measures. The common co-occurrence measures are defined as follows:

$$Dice(ij) = 2 * a / (f_i + f_j)$$

$$PMI(i,j) = \log a - (\log f_i + \log f_j)$$

$$MI(i,j) = \log a - \log (a + b) - \log (a + c)$$

$$LLR(i,j) = a \log a + b \log b + c \log c + d \log d - (a+b) \log (a+b) - (a+c) \log (a+c) - (b+d) \log (b+d) - (b+d) \log (c+d) + (a+b+c+d) \log (a + b + c + d)$$

$$\Phi^2 (i,j) = (ad - bc)^2 / ((a+b) (a+c) (b+c) (b+d)$$

In addition to the basic count frequencies and co-occurrence measures, more recent work on using statistical properties for multi-word term extraction focused on related frequencies such as the nested term frequency and the no. of tokens of a term candidate.

3.3. Statistical Properties

Distributional properties can be viewed as localized statistical properties. The statistical properties in the previous section make use of global count occurrences of words to extract co-occurrence statistics between words. The distributional properties relate to (i) the number of documents that a word occurs within a corpus and/or (ii) the differing counts of a word occurring across two or more corpora.

A common measure is the *term frequency – inverse document frequency* (*tf-idf*). The term frequency reflects the global counts of a word and the inverse document frequency measures the spread of the word throughout the document collection. Formally,

Let f_i be the no. of times a term occurs in all documents

Let n_i be the no. of documents where the term, i, occurs

Let N_{doc} be the total no. of documents in a corpus

The term frequency: $tf = f_i$

The document frequency: $df = n_i / N_{doc}$

Thus, inverse document frequency: $idf = N_{doc} / n_i$

And, $tf\text{-}idf = tf * idf$

A high word frequency might favor the global statistical co-occurrence measure however if the mass of the counts comes from a low number of documents, it will reflect a low *tf-idf* score deeming the term to be document-specific.

Other than using the distributional properties of words within a corpus, it is also helpful to compare the distribution of words across corpora. By comparing a domain specific corpus distribution to a general corpus, we can determine the weirdness of ratio of term frequencies across the corpora (Ahmad et al. 2007). The weirder a term, the more domain-specific a term is and the more likely it is to be a term candidate to form the terminology of a specific domain. We can simply refer to the relative frequency ratio across the corpora as such:

$$Weirdness = (f_i^D * N^G) / (f_i^G * N^D)$$

where f_i^D *is the term frequency of the word, i in the domain-specific corpus* D *and* f_i^G *is the term frequency of in the general corpus* G.

3.4. C-Value and NC-Value

Frantzi et al. (1998; 2000) introduced a method to use both linguistic and statistical information using C-value and NC-value. They start with a set of POS patterns and a stop word list to pre-filter possible n-grams before they calculate the n-gram's termhood using the C-value metric and the concept of nested terms. Nested terms refer to those terms that appear within other longer terms and may or may not appear by themselves in the corpus (Frantzi et al. 1998), e.g. '*floating point*' is a nested term because it is also found in '*floating point arithmetic*'.

For non-nested terms, the C-value accounts for the length of the term candidate and its frequency. For nested terms, the C-value subtracts the average number of times the term is nested in other term n-grams. Thus if 'floating point' occurs as a nested term candidate as often or more than it does as an independent term, then it will have low C-value. Formally:

Let NG *be the set of all n-grams possible from a corpus.*

Let T *be the set of all n-grams possible after using a POS pattern filter, i.e.*

$$T \subset NG$$

Let t *be a candidate term that is filtered from the full list of n-grams, and*

T_N *be the set of terms that contains nested terms with* t *such that*

$t \in TN \in T$

Note that f_i refers to the frequency of a term, i, and $|t|$ referring to the no. of words in the term.

$$C\text{-}value(t) = \begin{cases} \log |t| * f_t & \text{if } t \text{ is not nested} \\ \log |t| * (f_t - (1 / f_{TN}) * \Sigma_{i \in TN} f_i) & \text{otherwise} \end{cases}$$

From the C-Value equation, the C-value will be high for long non-nested strings with high frequency. The limitation of the C-value is that it can only be applied to multi-word terms.

And extension of the C-value is the NC-value which accounts for the context which the term occurs. The NC-value re-ranks the term candidates extracted from the C-values by looking into the previous words occurring before the term. This is motivated by the notion of extended terms, where the terms constrain the modifiers they accept (Sager et al. 1980). This contextual constraint manifests itself as a weight to account for the number of nested terms within in the candidate term; it is then normalized by the cumulative context weight (CCW). Formally it is defined as follows:

Let f_{t_i} be the frequency of the word j appearing prior to term t

Let f_{n_i} be no. of terms that follows word j and n be the total no. of terms

$$CCW(j) = \Sigma f_{t\_j} * f_{n\_j} / n$$

$$NC\text{-}value(t) = 0.8\ C\text{-}value(t) + 0.2\ CCW(j)$$

C-values and NC-values have proved to perform well (Zhang et al. 2008). However it does not measure termhood as defined by Kageura and Umino (1996). The formulation of the NC-value measures how consistently a phrase can be a term but it does not exactly contribute to select any n-grams to be a term. Inherently, the term candidate selection is handled by the POS pattern filter and the NC-value re-ranks the terms to further threshold the list of candidates. Although it was a solution created close to a decade ago, it is still a common algorithm used for commercial and academic term extraction[4].

4. A Novel Approach: Querying a Language Model to Measure Termhood

The calculation for frequencies described so far in our brief term extraction survey is based on raw counts of a word or term. N-gram language models have developed and applied to other NLP applications such as speech processing and

[4] *https://code.google.com/p/jatetoolkit/wiki/JATEIntro*

machine translation (e.g. Kirchoff and Yang, 2005; Li and Khundapur, 2008; Schwenk et al. 2012; Lembersky et al. 2012; Chelba et al. 2012).

The major advantage of using a language model is the possibility of accounting for unknown words using interpolation and smoothing techniques (Chen and Goodman, 1996; Chelba et al. 2010). By using a language model, we avoid the need to optimize n-gram counting when implementing the term extraction algorithm, especially when very fast implementations of language models already exists (Heafield, 2011).

The Pointwise Mutual Information (PMI) of any term can be calculated with a backoff trigram language model as follows:

> *Let $PMI_{LM}(t)$ be the pointwise mutual information (PMI) score based on language model termhood probability of a term t occurring*
>
> *Let i be a possible n-gram in a term such that*
>
> $$i \in t \text{ and } t = \langle i_1, \ldots, i_n \rangle$$
>
> *where n is the total no. of possible n-grams, excluding unigrams, in t*
>
> *Let u be any unknown word not seen in the corpus, represented by <unk>*

$$PMI_{LM}(t) = 1/n \sum_{i \in t} \begin{cases} log\ PMI(i_1, i_2) & \text{if all words in term in found in training corpus} \\ PMI_{LM}(u, i) & \text{Otherwise, when } i \text{ is an unknown word} \end{cases}$$

The brevity normalization ($1/n$) does not favour a longer term compared to the C-value. The NC-value and the inner context probabilities have either been handled by the n-gram model during training or the backoff probability. When an unknown word exists in the term, it uses the backoff probability with the contextual probability of an unknown word occurring in the i_{-1} and i_{-2} context.

For example, for the word 'floating point arithmetic' and '*floating point routine'*, we assume that the word '*arithmetic'* is not found in the language model training corpus. Hence the other probabilities were calculated by the language model as follows:

$$\begin{aligned} PMI_{LM}(\text{'floating point routine'}) = 1/5\ (&log\ PMI_{LM}(\text{'', 'floating'}) \\ &+ log\ PMI_{LM}(\text{'floating', 'point'}) \\ &+ log\ PMI_{LM}(\text{'point', 'routine'}) \\ &+ log\ PMI_{LM}(\text{'floating point', 'routine'})) \\ &+ log\ PMI_{LM}(\text{'floating', 'point routine'})) \end{aligned}$$

$$\begin{aligned} PMI_{LM}(\text{'floating point arithmetic'}) = 1/5\ (&log\ PMI_{LM}(\text{'', 'floating'}) \\ &+ log\ PMI_{LM}(\text{'floating', 'point'}) \\ &+ log\ PMI_{LM}(\text{'point', '<unk>'}) \end{aligned}$$

$$+ \log PMI_{LM}(\text{'floating point'}, \text{'<unk>'}))$$
$$+ \log PMI_{LM}(\text{'floating'}, \text{'point <unk>'}))$$

where,

$$PMI_{LM}(\text{'floating'}, \text{'point'}) = \log P(\text{'floating point'}) - (\log P(\text{'floating'}) + \log(\text{'point'}))$$

$$PMI_{LM}(\text{''}, \text{'floating'}) = PMI_{LM}(\text{<S>}^{5}, \text{'point'})$$

$$PMI_{LM}(\text{'point arithmetic'}) = \log P(\text{'point <UNK>'}) - (\log P(\text{'point'}) + \log(\text{'unknown'}))$$

$$P(\text{'point <UNK>'}) = P(\text{'point'}, \text{<UNK>}) / P(\text{<UNK>})$$

Similar to the C-value and NC-value, our approach is limited to multi-word expressions. Our approach is (i) a more flexible approach than C-value that allows querying a smoothed language model to retrieve probabilities and (ii) relies on the backoff probabilities that have the ability to score unknown words. The traditional C-value would count co-occurrence instances without smoothed counts and would not handle unknown words.

We note that for the single term PMI, we force a PMI value to be calculated with the start of sentence symbol that resulted from the left and right sentence padding during the language model training. If we use the previous context word before the term, we could get a similar candidate re-ranking as from the NC-value. We did not experiment with the context because we note that finding the correct weights for the C-value and CCW should not be as trivial as assigning static 0.2 and 0.8 coefficients in the NC-value calculations. From here on, we refer to our term extraction metric described above as the Language-model PMI (LPMI).

4.1. Experimental Setup

We evaluate our novel approach on the food domain corpus (WikiFood) that was built for an ontology induction task at SemEval-2015 (Tan et al. 2015). The corpus contains 869 food terms and the relevant Wikipedia articles that contain these terms. Of those terms 752 terms are multi-words and from the 752 multi-words, we extracted 42,851 sentences (1,207,677 tokens) that contains the multi-word terms. We expect only 1 correct term to be extracted per sentence.

To access the probabilities for calculating LPMI, we trained a 5-gram language model on the corpus and we evaluate the LPMI accuracy against the traditional C-value score in extracting terms from the corpus. For each metric we extract the top five term candidates to match against the 1 correct term per sentence.

We evaluate the metrics by calculating the accuracy of the top ranked term candidates for each sentence and matching them against the correct term for that

[5] <S> refers to start of sentence.

sentence. Since the experimental task is structured more like an information retrieval task of candidate ranking, we use the mean reciprocal rank (MRR) to evaluate the ranking efficiency of the term extraction metrics. The mean reciprocal rank is calculated by averaging the ranks of the retrieved candidates against all possible candidates.

4.2. Results

	Accuracy (top 1)	**Accuracy (out of 5)**	**MRR**
LPMI	0.2829	0.4518	1.632
C-value	0.2326	0.3226	2.263

Table 1: Accuracy and Mean Reciprocal Rank for Term Extraction for WikiFood Corpus

Table 1 presents the results of the experiment on term extraction for the WikiFood Corpus. The LPMI metric clearly scores better in terms of accuracy to rank the correct term candidate in the top position with a mean rank of 1.63 and an accuracy of 0.28 compared to the C-value's mean rank of 2.26 and an accuracy of 0.23.

C-Value		**LPMI**	
Bull's-Eye Barbecue Sauce	-6.491	Burger King	-3.269
Al Steak Sauce	-7.078	Bull's-Eye Barbecue Sauce	-4.909
Kraft products	-7.397	A1 Steak Sauce	-6.216
Burger King	-8.512	Barbecue Sauce	-6.304
A1 Steak	-9.148	Kraft products	-6.675

Table 2: An Example Output of Ranked Term Candidates and Metrics Scores

However, we do note the low accuracy scores because a term candidate with high termhood might not necessarily be the query term that we are expecting from the sentence. For example in the sentence "*In both cases, Burger King prominently used the names of the Kraft products, A1 Steak Sauce and Bull's-Eye Barbecue Sauce, in the names of the sandwiches*". The LPMI and C-value extracted the following terms in Table 2.

Although the absolute score between the two metrics are on a different scale but are comparable because they are constricted by a probability in the logarithm space. We see that there are many valid terms extracted (e.g. "Burger King" and "Bull's-Eye Barbecue Sauce") by both metrics. However, they are not used in the accuracy computation because of the aim to build a food terminology for a taxonomy and to check against a gold standard terminology.

Unsurprisingly, the C-value favours longer terms and ranks them higher.
Additionally, the target "Barbecue Sauce" has been excluded in the top 5 terms for the C-value due to its preference for nested terms and that allowed "A1 Steak" into the top 5 C-value extracted terms.

4.3. Conclusion

We have developed a novel approach to terminology extraction using the Language-Mode PMI metric. Our preliminary experiments on extracting the food terminology from the WikiFood corpus showed similar or better results than the commonly used C-value. The main advantage our approach is the use of smoothen frequency from a pre-built language model that efficiently calculates the logarithmic probabilities and the ability to assign a probability to an unknown word, which was previously not possible.

5. Ontology

Traditional ontologies, such as CYC (Lenat, 1995) and SUMO (Niles and Pease, 2001), were manually crafted with much effort and suffer from coverage sparsity. This motivated the current researches towards automatic ontology induction from texts (Lin and Pantel, 2001; Snow et al. 2006; Velardi et al. 2013). Ontology induction research has moved from semi-manually crafted rule-based systems towards fully unsupervised clustering, graph-based and vector space approaches.
A variety of methods are used in taxonomy induction. They can be broadly categorized as (i) pattern/rule based, (ii) clustering based, (iii) graph-based and (iv) vector space approaches.

5.1. Pattern/Rule Based Approaches

Hearst (1992) first introduced ontology learning by exploiting lexico-syntactic patterns that explicitly link a hypernym to its hyponym, e.g. "*X and other Ys*" and "*Ys such as X*". These patterns could be manually constructed (Berland and Charniak, 1999; Kozareva et al., 2008) or automatically bootstrapped (Girju, 2003). These methods rely on surface-level patterns and incorrect items are frequently extracted because of parsing errors, polysemy, idiomatic expressions, etc.

5.2. Clustering Based Approaches

Clustering based approaches are mostly used to discover hypernym (*is-a*) and synonym (is-like) relations. For example, to induce synonyms, Lin (1998) clustered words based on the amount of information needed to state the commonality between two words.

Contrary to most bottom-up clustering approaches for taxonomy induction (Caraballo, 2001; Lin, 1998), Pantel and Ravichandran (2004) introduced a top-down approach, assigning the hypernyms to clusters using co-occurrence statistics and then pruning the cluster by recalculating the pairwise similarity between every hyponym pair within the cluster.

5.3. Graph Based Approaches

In graph theory (Biggs et al., 1976), similar ideas are conceived with a different jargon. In graph notation, nodes/vertices form the atomic units of the graph and nodes are connected by directed edges. A graph, unlike an ontology, regards the

hierarchical structure of a taxonomy as a by-product of the individual pairs of nodes connected by a directed edges. In this regard, a single root node is not guaranteed to produce a tree-like structure.

Disregarding the overall hierarchical structure, the crux of graph induction focuses on the different techniques of edge weighting between individual node pairs and graph pruning or edge collapsing (Kozareva and Hovy, 2010; Navigli et al., 2011; Fountain and Lapata, 2012; Tuan et al., 2014).

5.4. Vector Space Approaches

Semantic knowledge can be thought of as a two-dimensional vector space where each word is represented as a point and semantic association is indicated by word proximity. The vector space representation for each word is constructed from the distribution of words across context so that words with similar meaning are found close to each other in the space (Mitchell and Lapata, 2010; Tan, 2013).

Although vector space models have been used widely in other NLP tasks, ontology/taxonomy induction using vector space models has not been popular. It is only since the recent advancement in neural nets and word embeddings that vector space models are gaining ground for ontology induction and relation extraction (Saxe et al., 2013; Khashabi, 2013).

Most recently, Fu et al. (2014) discovered that hypernym-hyponyms pairs have similar semantic properties as the linguistics regularities discussed in Mikolov et al. (2013b). For example:

$$v(shrimp)\text{-}v(prawn) \approx v(fish)\text{-}v(goldfish)$$

Intuitively, the assumption is that all words can be projected to their hypernyms based on a transition matrix, i.e. given a word *x* and its hypernym y, a transition matrix Φ exists such that $y = \Phi x$, e.g.

$$v(goldfish) = \Phi \times v(fish)$$

Fu et al. proposed two projection approaches to identify hypernym-hyponym pairs:

- ***uniform linear projection*** where Φ is the same for all words and Φ is learnt by minimizing the mean squared error of Φ x-y across all word-pairs (i.e. a domain independent Φ) and
- ***piecewise linear projection*** that learns a separate projection for different word clusters; i.e. a domain dependent Φ, where a taxonomy's domain is bounded by its terms' cluster(s).

In both projections, hypernym-hyponym pairs are required to train the transition matrix Φ.

6. A Novel Approach: Querying a Language Model to Measure Termhood

Instead of learning a supervised transition matrix Φ, we propose a simpler unsupervised approach where we learn a vector for the phrase "*is-a*". We single-tokenize the adjacent "*is*" and "*a*" tokens and learn the word embeddings with *is-a* forming part of the vocabulary in the input matrix.

Effectively, we hypothesize that Φ can be replaced by the *is-a* vector. To achieve the piece-wise projection effects of Φ, we trained a different deep neural net model for a specific domain and assume that the *is-a* scales automatically with respect to the domain-specific corpus it is trained on.

For example, in the food domain, we learn the neural net vector for *is-a* from a food domain specific corpus, and we hypothesize that the multiplication of the $v(tiramisu)$ and the $v(is\text{-}a_{food})$ vectors yields a proxy vector and we consider the top 10 word vectors that are most similar to this proxy vector as the possible hypernyms, formally:

Let F be the set of all possible terms within the food domain such that

$$F = \langle f_1, \ldots, tiramisu, \ldots, cake, f_n \rangle \text{ and}$$

$$v(tiramisu) \times v(is\text{-}a_{food}) \equiv v(proxy), \text{ where } proxy \notin F$$

$$v(proxy) \approx v(cake)$$

$$v(cake) \in \operatorname{argmax} v(proxy) \cdot v(f), \text{ where } f \in F, \; n=10$$

6.1. Experimental Setup

We experimented with our novel approach for ontology induction and evaluated against the SemEval-2015 taxonomy induction shared task and achieved competitive performance as compared to other state-of-art systems that participated in the shared task. The rest of this section will describe the experiment setup, the evaluation criteria and the results of the system's performance.

Similar to Fountain and Lapata (2012), the SemEval-2015 Taxonomy Extraction Evaluation (TaxEval) task addresses taxonomy learning without the term discovery step (Bordea et al., 2015). The focus is on creating the hypernym-hyponym relations. In the TaxEval task, taxonomies are evaluated through comparison with gold standard taxonomies.

The four different domains for the shared task were chemicals, equipment, food and science. The gold standards used in evaluation are the ChEBI ontology for the chemical domain (Degtyarenko et al., 2008), the *Material Handling Equipment taxonomy*[6] for the equipment domain, the *Google product taxonomy*[7] for the food

[6] *http://www.ise.ncsu.edu/kay/mhetax/index.htm*
[7] *http://www.google.com/basepages/producttype/taxonomy.en.txt*

domain and the *Taxonomy of Fields and their Different Sub-fields*[8] for the science domain. Furthermore, all four domains are also evaluated against the sub-hierarchies from the WordNet ontology that subsumes the Suggested Upper Merged Ontology (Pease et al. 2002).

To produce a domain specific corpus for each of the given domains in the task, we used the Wikipedia dump and pre-processed it using WikiExtractor[9], we then extracted documents that contain the terms for each domain individually. We trained a skip-gram model phrasal word2vec neural net (Mikolov et al., 2013a) using gensim (Rehurek and Sojka, 2010). We trained the neural nets for 100 epochs with a window size of 5 for all words in the corpus. The *is-a* phrase was single tokenized before the neural net model training.

The multi-faceted evaluation scheme presented in Navigli (2013) was adopted to compare the overall structure of the taxonomy against a gold standard, with an approach used for comparing hierarchical clusters. The multi-faceted evaluation scheme is defined (i) the structural measures of the induced taxonomy and the comparison against gold standard taxonomy.

6.2. Results

	|V|	**|E|**	**#c.c**	**cycles**	**#VC**	**%VC**	**#EC**	**%EC**	**:NE**
Chemical	13785	30392	302	YES	13784	**0.784**	2427	0.098	1.127
Equipment	337	548	28	YES	336	0.549	227	0.369	0.522
Food	1118	2692	23	YES	948	0.609	428	0.270	1.427
Science	335	952	14	YES	354	0.783	173	**0.372**	1.675
WN Chemical	1173	3107	31	YES	1172	**0.868**	532	**0.384**	1.857
WN Equipment	354	547	43	YES	353	0.743	149	0.307	0.821
WN Food	1200	3465	23	YES	1199	0.807	549	0.358	1.902
WN Science	307	892	8	YES	306	0.713	156	0.355	1.669

Table 1: Structural Measures and Comparison against Gold Standards

Table 1 summarizes the preliminary results we achieved for our ontology induction system. The left columns of the table describe the structural integrity of the ontology induced (i.e. the "tree-like-ness" of the ontology) and the right columns of the table presents the ontology comparison against the gold standards (i.e. the "correct-ness" of the onotology).

The labels of the columns refer to no. of distinct vertices and edges in induced taxonomy (**|V|** and **|E|**), no. of connected components (**#c.c**), whether the taxonomy is a Directed Acyclic Graph (**cycles**), vertex and edge coverage, i.e. proportion of gold standard vertices and edges covered by system (**%VC** and **%EC**), no. of vertices and edges in common with gold standard (**#VC** and **#EC**) and ratio of novel edges (**:NE**).

[8] *http://sites.nationalacademies.org/PGA/Resdoc/PGA044522*
[9] *http://medialab.di.unipi.it/wiki/Wikipedia_Extractor*

In terms of vertex coverage, our system performs best in the chemical and WordNet chemical domains. Regarding edge coverage, our system achieves highest coverage for the science domain and WordNet chemical domain. Having high edge and vertex coverage significantly lowers false positive rate when evaluating hypernym-hyponyms pairs with precision, recall and F-score.

We also note that the Wikipedia corpus that we used to induce the vectors lacks coverage for the food domain. In the other domains, we discovered all terms in the Wikipedia corpus plus the domains' root hypernym (i.e. |V| = #VC + 1). In the food domain however, we only managed to cover 948 out of the 1118 terms.

	Chemical	Equipment	Food	Science	WN Chemical	WN Equipment	WN Food	WN Science
P	0.080	0.414	0.159	0.182	0.171	0.272	0.158	0.175
R	0.098	0.369	0.270	0.372	0.384	0.307	0.358	0.354
F	0.088	0.390	0.200	0.244	0.237	0.289	0.220	0.234
PNE	-	80%	34%	34%	-	45%	25%	34%

Table 2: Precision (P), Recall (R) and F-score (F) of the Induced Edges and the Precision of the Novel Edges (PNE)

Table 2 presents the performance of our ontology induction approach in classic precision, recall and F-score for the edges induced by the system. We achieved high recall in terms of vertices coverage for the chemical domain. However, our edge coverage is low, thus it naturally leads to the lowest F-score among the domains evaluated.

Although we were unable to capture all the vertices in the food domain due to the corpus coverage, we see that the system performs reasonably well in its F-score. Table 2 shows that our highest performs come from the equipment domain and the consistent recall and precision scores. Compared against the other domains, the equipment domain has the lowest novel edge ratio and hence the consistent F-score is expected.

The last row of Table 2 presents the manual evaluation of the novel edges precision, the novel edges were checked manually for the equipment, food and science domains. The chemical domain was excluded due to the lack of domain experts in the evaluation process. In general, 1 in 3 novel edges produced by our systems in the food and the science domain are accurate and 4 in 5 of the novel edges in the equipment domain are correct. This shows the potential of our approach to not only produce ontology similar to the gold standards but also provide ontological edges that the gold standard missed.

6.3. Results

In fulfilment of the WP4.2 deliverable, we developed a novel unsupervised approach to ontology induction using the *is-a* vector learnt from a neural net model of the Wikipedia corpus. Our preliminary experiments on the SemEval-2015 Taxonomy Induction task dataset shows promising results. Given the simple approach to hypernym-hyponym relations, it is possible that future research can apply the method to other non-content words vectors to induce other relations between entities.

7. Ontology

In both parts of this document, we summarized our findings for the novel approaches in terminology extraction and ontology induction.

For terminology, we introduced the Language-model PMI (LPMI) association measure that makes use of pre-computed language model. We showed that it performs better than the commonly used C-value measure. The use of language models in terminology extraction brings the terminology research closer to machine translation research that depends on language models for decoding the translation outputs. Previous researches have shown that using an extracted terminology in machine translation improves the translation quality (Tsvetkov and Wintner, 2012; Simova and Kordoni, 2013; Tan and Pal; 2014). Currently, we are working on using LPMI extracted terms in machine translation experiments and we expect to see improvement in translation quality.

For ontology, we introduce a novel approach that learns a feature function word embedding vector from the is-a phrase to induce a hypernym vector by extrapolating the hyponym vector. The approach showed promising results in the SemEval-2015 Taxonomy Induction task dataset. Although incorporating ontological knowledge in statistical machine translation is a challenging (Knight, 1993; Knight and Luk, 1994; Skadins 2011), recent advancement in neural network machine translation (Devlin et al. 2014; Wang et al. 2014) has made it easier to incorporate semantic information into machine translation. We foresee the integration of our word embedding approach to ontology induction into neural network machine translation.

References

- Khurshid Ahmad, Lee Gillam, and Lena Tostevin. 1999. University of Surrey Participation in TREC8: Weirdness Indexing for Logical Document Extrapolation and Retrieval (WILDER). In Proceedings of TREC8.
- Matthew Berland and Eugene Charniak. 1999. Finding Parts in Very Large Corpora. In Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, pages 57–64.
- Norman Biggs, E. Keith Lloyd, and Robin J. Wilson. 1976. Graph theory 1736-1936. Clarendon Press.
- Georgeta Bordea, Paul Buitelaar, Stefano Faralli, and Roberto Navigli. 2015. Semeval-2015 task 17: Tax-onomy Extraction Evaluation. In Proceedings of the 9th International Workshop on Semantic Evaluation. Association for Computational Linguistics.
- Sharon Ann Caraballo. 2001. Automatic Construction of a Hypernym-labeled Noun Hierarchy from Text. Ph.D. thesis, Providence, RI, USA. AAI3006696.
- Ciprian Chelba, Dan Bikel, Maria Shugrina, Patrick Nguyen, and Shankar Kumar. 2012. Large scale language modeling in automatic speech recognition. arXiv preprint arXiv:1210.8440.
- ClientSide News. 2006. A New Mid-level Solution for Terminology Management. Retrieved from http://www.lexicool.com/lingo4-terminology-management-article.asp?IL=1
- Béatrice Daille. Study and implementation of combined techniques for automatic extraction of terminology. 1996. The balancing act: Combining symbolic and statistical approaches to language 1. pages 49-66.
- Ioo Dagan, Kenneth W. Church and William A. Gale. 1993. Robust bilingual word alignment for machine aided translation. In Proceedings of the Workshop on Very Large Corpora: Academic and Industrial Perspectives, 1-8, Columbus, Ohio.
- Kirill Degtyarenko, Paula De Matos, Marcus Ennis, Janna Hastings, Martin Zbinden, Alan Mcnaught, Rafael Alcantara, Michael Darsow, Micka¨el Guedj, and Michael Ashburner. 2008. ChEBI: A Database and Ontology for Chemical Entities of Biological In- terest. Nucleic acids research, 36(suppl 1):D344–D350.
- Jacob Devlin, Rabih Zbib, Zhongqiang Huang, Thomas Lamar, Richard Schwartz, and John Makhoul. 2014. Fast and robust neural network joint models for statistical machine translation. In 52nd Annual Meeting of the Association for Computational Linguistics, Baltimore, MD, USA.
- Guy Emerson, Liling Tan, Susanne Fertmann, Alexis Palmer, and Michaela Regneri. 2014. SeedLing: Building and Using a Seed corpus for the Human Language Project. In Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages, pages 77–85.
- Helmut Felber. 1984. Terminology Manual. International Information Centre for Terminology (Infoterm).
- Trevor Fountain and Mirella Lapata. 2012. Taxonomy Induction using Hierarchical Random Graphs. In Proceedings of the 2012 Conference of the North American

Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 466–476.

- Katerina T. Frantzi, Sophia Ananiadou, and Junichi Tsujii. 1998. The c-value/nc-value method of automatic recognition for multi-word terms. Research and advanced technology for digital libraries. Springer Berlin Heidelberg, pages 585-604.
- Katerina Frantzi, Sophia Ananiadou, and Hideki Mima. 2000. "Automatic recognition of multi-word terms:. the c-value/nc-value method." International Journal on Digital Libraries 3.2. pages 115-130.
- Ruiji Fu, Jiang Guo, Bing Qin, Wanxiang Che, Haifeng Wang, and Ting Liu. 2014. Learning Semantic Hierarchies via Word Embeddings. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1199–1209.
- Xavier Gomez Guinovart and Alberto Simões. 2009. Parallel corpus-based bilingual terminology extraction. In Marie-Claude L'Homme and Sylvie Szulman, editors, 8th International Conference on Terminology and Artificial Intelligence, Toulouse, France.
- Roxana Girju. 2003. Automatic Detection of Causal Relations for Question Answering. In Proceedings of the ACL 2003 workshop on Multilingual summarization and question answering-Volume 12, pages 76–83.
- Kenneth Heafield. 2011. KenLM: Faster and Smaller Language Model Queries. In Workshop for Statistical Machine Translation (WMT) at EMNLP, Edinburgh, Scotland, United Kingdom.
- Marti A Hearst. 1992. Automatic Acquisition of Hyponyms from Large Text Corpora. In Proceedings of the 14th conference on Computational linguistics Volume 2, pages 539–545.
- Andrew Hippisley, David Cheng, and Khurshid Ahmad. 2005. The head-modifier principle and multilingual term extraction. Natural Language Engineering 11.02. pp. 129-157.
- Ana Hoffmeister. 2014. Terminology Processes and Quality Assurance [slides]. Presented in Terminology for Translators Colloquium. Saarland, Germany. Retrieved from http://fr46.uni-saarland.de/fileadmin/user_upload/personen/wurm/Workshops/Hoffmeister_Terminology_Processes_and_Quality_Assurance.pdf.
- John S. Justeson and Slava M. Katz. 1995. Technical terminology: some linguistic properties and an algorithm for identification in text. Natural language engineering 1.01. pages 9-27.
- Kyo Kageura and Bin Umino. 1996. Methods of automatic term recognition: A review. Terminology 3, no. 2. pp. 259-289.
- Daniel Khashabi. 2013. On the Recursive Neural Networks for Relation Extraction and Entity Recognition. Technical report.
- Katrin Kirchhoff, and Mei Yang. 2005. Improved language modeling for statistical machine translation. In Proceedings of the ACL Workshop on Building and Using Parallel Texts, pp. 125-128. Association for Computational Linguistics.
- Kevin Knight. 1993. Building a large ontology for machine translation. In Proceedings of the workshop on Human Language Technology, pp. 185-190.

- Kevin Knight and Steve K. Luk. 1994. Building a large-scale knowledge base for machine translation. In AAAI, vol. 94, pp. 773-778.
- Zornitsa Kozareva and Eduard Hovy. 2010. A Semi-Supervised Method to Learn and Construct Taxonomies using the Web. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pages 1110–1118.
- Zornitsa Kozareva, Ellen Riloff, and Eduard Hovy. 2008. Semantic Class Learning from the Web with Hyponym Pattern Linkage Graphs. In Proceedings of ACL-08:HLT, pages 1048–1056, Columbus, Ohio, June.
- Julian Kupiec. 1993. An algorithm for finding noun phrase correspondences in bilingual corpora. In Proceedings of the 31st Annual Conference of the Association for Computational Linguistics, 17-22, Columbus, Ohio.
- Gennadi Lembersky, Noam Ordan, and Shuly Wintner. 2012. Language models for machine translation: Original vs. translated texts. Computational Linguistics 38, no. 4. pp. 799-825.
- Douglas B Lenat. 1995. CYC: A Large-Scale Investment in Knowledge Infrastructure. Communications of the ACM, 38(11):33–38.
- LZhifei Li, and Sanjeev Khudanpur. 2008. Large-scale discriminative n-gram language models for statistical machine translation. In Proceedings of AMTA, pp. 133-142.
- Dekang Lin. 1998. Automatic Retrieval and Clustering of Similar Words. In Proceedings of the 17th international conference on Computational linguistics Volume 2, pages 768–774.
- Dekang Lin and Patrick Pantel. 2001. Discovery of Inference Rules for Question-Answering. Natural Language Engineering, 7(04):343–360.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.
- Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013b. Linguistic Regularities in Continuous Space Word Representations. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 746–751.
- George A Miller. 1995. WordNet: A Lexical Database for English. Communications of the ACM, 38(11):39–41.
- Jeff Mitchell and Mirella Lapata. 2010. Composition in Distributional Models of Semantics. Cognitive Science, 34(8):1388–1439.
- Roberto Navigli, Paola Velardi, and Stefano Faralli. 2011. A Graph-Based Algorithm for Inducing Lexical Taxonomies from Scratch. In IJCAI 2011, Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain, July 16-22, 2011, pages 1872–1877.
- Patrick Pantel and Deepak Ravichandran. 2004. Automatically Labeling Semantic Classes. In Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004.
- Adam Pease, Ian Niles, and John Li. 2002. The Suggested Upper Merged Ontology: A Large Ontology for the Semantic Web and its Applications. In Working Notes of the AAAI-2002 Workshop on Ontologies and the SemanticWeb, Edmonton, Canada.

- Jörg Porsiel. 2008. Machine translation at Volkswagen: a case study. Multilingual Computing & Technology, vol, 100.
- Jörg Porsiel. 2011. Machine translation at Volkswagen [slides]. Presented in Third Joint EM+/CNGL Workshop. Retrieved from http://mtmarathon2010.info/JEC2011_Porsiel _slides.pdf
- Radim Rehurek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In Proceedings of the LREC 2010 Workshop on New Chal-lenges for NLP Frameworks, pages 45–50, Valletta, Malta, May. ELRA.
- Luıs Sarmento, Paula Carvalho, and Eugenio Oliveira. 2009. Exploring the Vector Space Model for Finding Verb Synonyms in Portuguese. In Proceedings of the International Conference RANLP-2009, pages 393–398.
- Andrew M. Saxe, James L. McClelland, and Surya Ganguli. 2013. Learning Hierarchical Category Structure in Deep Neural Networks. pages 1271–1276.
- Holger Schwenk, Anthony Rousseau, and Mohammed Attik. 2012. Large, pruned or continuous space language models on a gpu for statistical machine translation. In Proceedings of the NAACL-HLT 2012 Workshop: Will We Ever Really Replace the N-gram Model? On the Future of Language Modeling for HLT, pp. 11-19. Association for Computational Linguistics.
- Iliana Simova and Valia Kordoni. 2013. Improving English-Bulgarian statistical machine translation by phrasal verb treatment. In Proceedings of MT Summit XIV Workshop on Multi-word Units in Machine Translation and Translation Technology, Nice, France.
- Raivis Skadins. 2011. Spatial Ontology in Factored Statistical Machine Translation. In Proceedings of the 2011 conference on Databases and Information Systems VI: Selected Papers from the Ninth International Baltic Conference, DB&IS 2010, pp. 153-166. IOS Press.
- Frank Smadja and Kathleen McKeown. 1994. Translating collocations for use in bilingual lexicons. In Proceedings of the ARPA Human Language Technology Workshop 94, Plainsboro, New Jersey.
- Rion Snow, Daniel Jurafsky, and Andrew Y Ng. 2006. Semantic Taxonomy Induction from Heterogenous Evidence. In Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, pages 801–808.
- Liling Tan. 2013. Examining crosslingual word sense disambiguation. Master's thesis, Nanyang Technological University. pages 17-21.
- Liling Tan, Rohit Gupta and Josef van Genabith. 2015. USAAR-WLV: Hypernym Generation with Deep Neural Nets. In Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015). Denver, Colorado.
- Liling Tan and Santanu Pal. 2014. Manawi: using multi-word expressions and named entities to improve machine translation. In Proceedings of Ninth Workshop on Statistical Machine Translation. Baltimore, USA.
- Stefan Thater, Hagen F¨urstenau, and Manfred Pinkal. 2010. Contextualizing Semantic Representations using Syntactically Enriched Vector Models. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 948–957.

- Yulia Tsvetkov and Shuly Wintner. 2012. Extraction of multi-word expressions from small parallel corpora. Natural Language Engineering 18, no. 04 (2012): 549-573.
- Luu Anh Tuan, Jung-jae Kim, and Kiong See Ng. 2014. Taxonomy construction using syntactic contextual evidence. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 810–819.
- Kara Warburton. 2005. Terminology: Getting Down to Business, The Globalization Insider. Retrieved from http://web.archive.org/web/20110523103011/ http://www.lisa.org/globalizationinsider/2005/07/terminology_get.html
- Paola Velardi, Stefano Faralli, and Roberto Navigli. 2013. OntoLearn Reloaded: A Graph-based Algorithm for Taxonomy Induction. Computational Linguistics, 39(3):665–707.
- Rui Wang, Hai Zhao, Bao-Liang Lu, Masao Utiyama, and Eiichro Sumita. 2014. Neural Network Based Bilingual Language Model Growing for Statistical Machine Translation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP-2014): 189--195, Doha, Qatar.
- Eugen Wüster. 1969. Internationale Sprachnormung in der Technik, besonderes in der Elektrotechnik. Berlin: VDI, 1931 XV + 431 S. 2. Auflage. Bonn: Bouvier, 1966.
- Yongzheng Zhang, Evangelos Milios, and Nur Zincir-Heywood. 2004. A Comparison of Keyword-and Keyterm-Based Methods for Automatic Web Site Summarization. AAAI04 Workshop on Adaptive Text Extraction and Mining.